IEEE Big Data 2017 Panel Discussion on Bias and Transparency

Abhinav Maurya (Carnegie Mellon University;)
DOI: 10.1145/3236644.3236649

Panelists: Cynthia Dwork (Harvard University), **John Langford** (Microsoft Research), **Jure Leskovec** (Stanford University/Pinterest), **Jeanna Matthews** (Clarkson University)
Moderator: Ricardo Baeza-Yates (NTENT)
Scribe: Abhinav Maurya (Carnegie Mellon University)

In January 2017, the ACM US Public Policy Council released a report on algorithmic transparency and accountability (ACM US Public Policy Council, 2017) which outlined several characteristics for algorithms to be considered transparent and accountable:

- Awareness
- Access and redress
- Accountability
- Explanation
- Data Provenance
- Auditability
- Validation and Testing

A panel discussion on *Big Data Bias and Transparency* was organized at the IEEE International Conference on Big Data held in December 2017 to discuss opportunities and challenges faced by the data science community in their effort to incorporate the tenets of fairness, accountability, and transparency in their data-driven analyses and products. The panel consisted of Cynthia Dwork from Harvard University, John Langford from Microsoft Research, Jure Leskovec from Stanford University/Pinterest, Jeanna Matthews from Clarkson University, and was moderated by Ricardo Baeza-Yates from NTENT. This article provides an account of the panel discussion in the hope that it will be of interest to readers of AI Matters.

Ricardo Baeza-Yates: Perhaps the ideal form of interpretability is to have algorithms explain their decisions. Is it possible to build algorithms that can explain their decisions to convince us of their correctness?

Cynthia Dwork: I think it's difficult to pin down what a convincing explanation of a decision might be to a human. "Why was I turned down for the loan?" I have no idea how to answer that. There is a classifier, we feed your data in, this was the outcome, you were approved or turned down. A different question that I might be able to make sense of mathematically is "What is it that I can change at a reasonable cost that would lead to a different decision?" That is a question that makes some sense to me. But why was I classified this way, I can't really make sense out of it.

Jeanna Matthews: When we are talking about really important decisions like whether to send someone to jail or not, explanation might be even more important than an incredibly accurate learning algorithm. The ability to export human-readable, understandable version of important decisions makes them regulatable.

Jure Leskovec: As a community, we like taking datasets and training algorithms on them and competing on who gets the biggest AUC or F1 score. As we start thinking about really applying these methods to problems that have more consequence than whether you will click a given ad or not or maybe you will watch a given movie or not and maybe that ruins your Friday night, but that is the most serious consequence it has. When you start thinking about these more important societal applications, then the question becomes how humans and algorithms work together, and what kind of algorithms work with humans in a given way. So I think it's a much broader question than about just machine learning algorithms or systems.

John Langford: If you want your machine learning systems to be debuggable, you need to think about your model in the context of the data source. If you keep the model separate from the data source, that's a bug waiting to

happen. You need the data source attached to the model so that you can track back and discover why the model is beginning to behave in a certain way. Machine learning systems are more than just algorithms, they keep track of where the data is coming from and how it is used in training the model. For explanation, the same thing comes up. If you are trying to figure out how to create a better model, you need to explain its decisions and mistakes. Within every machine learning algorithm, there is always a bug. It is never the case that you have an optimal machine learning algorithm. There's always something you can do to improve it. Figuring why a machine learning algorithm is failing on a certain task is a great way to figure out how to improve it. So if you have ways of explaining why the model is behaving in a certain way or if you are operating in spaces where it is really obvious what the model is doing, these are mechanisms to figure out improvements to the model. Improvements to the ImageNet systems have been driven by figuring out what the bugs were in previous year's system and how to improve it. For auditability, trying to debug a non-deterministic system is really difficult. For accountability, we have a paper at the FAT-ML workshop showing that any cost-sensitive learning algorithm can be turned into a fairness-satisfying learning algorithm. Hence, if we know that there is a bias issue, we can modify our learning algorithms to address this issue systematically, at least in a much more systematic fashion than we do now.

Cynthia Dwork: (To Langford) What's your definition of fairness? Are they group definitions of fairness, or statistical parity, or something else...?

John Langford: There are several definitions. For every single definition, we can create a fairness-aware learning algorithm.

Cynthia Dwork: There are many fairness conditions that are mutually inconsistent. What happens in that case?

John Langford: What the definition of fairness should be is something that people need to figure out. But if you write down a definition of fairness and you want to have it forced, we can turn almost all definitions into a reduction which will transform a classifier into a fair learning algorithm.

Cynthia Dwork: So again I don't see how that can happen when these definitions are mutually exclusive.

John Langford: So you have to choose one. Given that you choose one, I will create a learning algorithm to give you what you want. If you don't choose one, then I can't do it.

Cynthia Dwork: And can you do it for individual fairness?

John Langford: What is the definition of individual fairness?

Cynthia Dwork: That similar people should be treated similarly. So you have some kind of metric for a given classification task which tells you how similar or dissimilar a given pair of people is. For this particular classification task, can you ensure that there is some relationship between the training distance and the probability distributions on their outcomes?

John Langford: Is this similar to equalized odds?

Cynthia Dwork: No, equalized odds is a group definition which says that this group as a whole should have similar outcome probabilities compared to other groups. But that's a group definition of fairness.

John Langford: I need to know the exact definition before I can give you an exact answer.

Ricardo Baeza-Yates: I think it might be useful to clarify what fairness is. Because politicians creating laws are unsure about it. One way to think about it is that when politicians create laws, they don't worry about the details of implementing the law. Some formalization of fairness can help politicians create rules and guidelines for programmers and businesses. My next question is on accountability of algorithms. Who is accountable for the transparency and fairness of learning algorithms? Is it the person providing the data, is it the one that programs the algorithm, is it the corporation which deploys the algorithm...? There are many implications for the future that will change the field. John, we start with you.

John Langford: The question of ethics in algorithms is related to fairness. We know that given a definition of fairness, you can train

any classification algorithm to comply with that definition, and tradeoff between accuracy and fairness. The problem is that many people don't have an actual definition of fairness or what is ethical. The second is that the algorithms have to be aware of protected attributes in order to be able to incorporate them for achieving fairness.

Ricardo Baeza-Yates: Jure, can you talk about your research with judges in this regard?

Jure Leskovec: Sure! I think these issues are really interesting and important. In our group, we have been working with the Chicago crime lab and with an economist here at Harvard. And the question we have been looking at is whether we can help criminal court judges make better bail decisions. The question is after a person is arrested, where will the person wait for trail. The person can await trial in jail, or they can be free. If free, they can misbehave - they can commit a violent crime, or they can commit a non-violent crime, or they can simply fail to appear at trial. So we were asking how can machine learning help judges make better bail decisions. It was interesting how many technical and algorithmic issues came out when we started working on this problem. Ricardo was saying that the law is very clear. And the law is that the judge should ignore the severity of the crime when making bail decisions. The judge should try to assess the probability of recidivism. As we were doing this research, one thing for example that we noticed was that machine learning algorithms could reduce the level of crime by around 40% if you keep the prison population the same. Another way to say this is if you keep the current level of crime, you could release 72-73% of the people awaiting bail. But the data collection process is itself biased. We only see the outcome of the people that were actually released, we don't know the outcomes for people who were kept in jail. If you assume that machine and human have access to the same information, there are statistical ways of imputation to get around this. But humans see much more than machines. To give an example of how bad this can be, consider that a judge learns from years of experience on sentencing young people that if their family shows up at the bail hearing, it is ok to release on bail. If the family doesn't show up, it means the person might commit another crime. Assume that we didn't go and encode this feature into our algorithm. Then, based on your release data, you will learn than young people who are released commit no crime. And any fancy or "fair" algorithm with the most proper cross-validation will tell you that young people commit no crime. And then you go and deploy this in the real world and your crime rates will go up. You suddenly see cases of young people committing crimes. What is interesting about this research is that even though the law is very clear, we developed a way to diagnose how humans make decisions and identified groups of defendants where judges make consistent decisions and on certain others they do not. For example, on single people who move around a lot and on families without kids, judges make very accurate decisions, but on people with kids, their decisions are much less accurate. One way to explain it is that this is because judges are making mistakes. Another way is to acknowledge that the objective/cost function of the judge and the algorithm are very different. If you put a single person in jail, you restrict their movement and ability to commit crime but there is little cost to the rest of the society. But if you put a person with kids in jail, there is a huge consequence on society, for the families and so on. And your decisions will now affect their future behavior. To make my long story short, very interesting things start happening if you take a real example and ask how can we build algorithms to help society and interesting aspects begin to emerge that one wouldn't even think about. And it's a very interesting area to work on.

Jeanna Matthews: So the question about ethics and societal values, those can change a lot with countries and even with different times in history. For example, in this country, we would rather let guilty people go free than put innocent people in jail. Another example is that one is not allowed to consider race in a hiring decision. I am very concerned that we are fundamentally replacing such societal values without any discussion of it. If you replace decisions like that with a piece of software, we run into problems. I think it is being labeled this way: these are unbiased logical decisions made by computers when that's not true. They are trained on historical data in which there is actually a lot of bias. Train-

ing on historical data makes sense; the past is all we have; we don't really have a choice. But we don't always want the decisions to look like the past. And when we fail to recognize it and think of it as a limitation of the tool, that's a problem. Another problem is the fundamental nature of these black-box algorithms that are being used. There are many proxies for the sensitive attributes such as race or gender. You could say that you are not looking at the sensitive attributes of race or gender, but proxies to those are what you could be looking at if you looked at the explanation. So would it be easy to build a software system that you could claim makes completely unbiased decisions and could keep certain people out of this country? Yes, you could do that and it might even be attractive in this political climate. That is very concerning. So if we have what we believe to be fundamental societal values and we are replacing human decision-making by black-box algorithmic decisions and we fail to require explanation of the decisions, we may be using sensitive attributes directly or using certain proxies which are just as good. Or it could be just a bug in the system that could lead it to make wrong decisions. In many of these systems, there is no forcing function for debugging. If you take a proprietary software system that you use to judge recidivism and the company says our intellectual property rights are more important than a defendant's right to explanation, does that sound outrageous to you? That's exactly what's happening, Louis versus Wisconsin for example. These are the things that are happening right now! We might be fundamentally changing our societal values without discussion simply by replacing human decisions with black-box decisions and without requiring explanations. I think there's a lot we can do in the technical community if we are sufficiently humble about limitations of the things we build and sufficiently advertise the dangers of using them and ways in which they are inappropriate. We should sound the alarm that they should not be used in ways that some people might want by tucking the details under the covers, sound the alarm on things that they would like to have happen by hiding under the label of completely unbiased decisions made by computers. At least, we can audit algorithms. It's more difficult to audit humans. So we have potential to do better. But we also have the potential to do a lot worse and label it as better.

Cynthia Dwork: Actually, I want to comment on a couple of things I heard before providing my answer. Thinking about fairness and the predictors that are used in legal contexts, I am not a lawyer or legal expert but I had a conversation with a PhD in Law grad student at Harvard. She was talking to me about bail decisions in New York state where the only factor one is allowed to take into account is flight risk. Ok, that's pretty concrete. But then she pointed out that there really are multiple reasons why one might be a flight risk. One is that they really might be a flight risk i.e. they might run away and not come to trial. But another is that they simply can't afford the transportation to get back to court. And incarcerating someone just because they cannot afford to get back to court seems wildly unfair. So these things are incredibly subtle and incredibly laden with context. Another thing that comes up in recidivism prediction. Imprisoning someone isn't only a question of protecting society from recidivists. There are other reasons for jailing people including punishment. Where does that get put into the mix? Having an estimate or a way of trying to estimate somehow the likelihood that somebody is going to do something violent is clearly useful but it's definitely not going to be the whole story. So this means that in order to decide sentencing, one has to sit down and decide what's the point of sentencing, and it involves enormous amount of societal context. About explanation, one of the things I hear already on the panel is these two different notions of explainability. One is explaining a particular decision, and another is akin to what you (Langford) were saying about debuggability. I dream about being able someday to say if we use this learning algorithm and these notions of fairness which I can lay out and you can examine and decide whether you like them or not, and these software principles for building systems that are fair; then maybe we could have something that is fair. I want us to get into that realm of things. We need ways that are much more systematic and catch issues besides the ones that we are already looking for. Fairness behaves oddly under composition. It does not behave like composition in cryptography or privacy-preserving data analysis. You can take two

things that are fair and you can find scenarios in which they are competing with each other, and the outcome of the system as a whole is not fair. So when we come back to me again, I will tell you a story about that.

Ricardo Baeza-Yates: The blackbox systems we deal with are so complex, and if we want to change how the system behaves, we need to understand the dynamics of the system. Also, there is the feedback loop. We collect data, make decisions, which changes the data we collect, changes the system, and so on. And the sideffects of these complex decisions cannot always be anticipated. For example, sending someone to jail might be the best way to turn someone into a criminal. You have the best training school and the best networking. So you are also changing probabilities of committing crimes in the future.

Jure Leskovec: I think this notion of exploration... If you think of bail or something else like medical procedures and so on, you cannot go ahead and collect random data. So if I want to build a skin cancer prediction system, the only way for me to collect data is for me to get a scalpel and start collecting samples, which would be amazingly non-ethical to do. I cannot come and start fooling around...

John Langford: I disagree. They do clinical trials all the time.

Jure Leskovec: No. But the point is they stop the clinical trial as soon as they have the result or they determine it is unethical. Clinical trials are not there to collect data; they are there to answer a specific question. And that's a huge difference. You cannot do random exploration. You cannot say: oh! we don't know what's happening here. Let's release this person.

John Langford: Random exploration need not be uniform or uninformed exploration. Uniform exploration is never the best kind of exploration.

Jure Leskovec: Even if it's non-random, I would say there are ethical issues with doing something that may be potentially harmful with the goal of collecting data.

Cynthia Dwork: So just to clarify, you are talking about collecting data to go into your training set? (Leskovec confirms.)

John Langford: I agree that there can be ethical issues, but I don't agree that every time you do exploration in the medical field, it is unethical. And clinical trials are a good example of this.

Jure Leskovec: Again, my point is medical trials are there to test hypotheses, not to collect data. The other thing that becomes interesting is the question of features. In bail, you have protected attributes like gender, religion, race, etc. On second thoughts, I think gender you can use but you can't use race or religion. Now, I think here's a good question that I don't have an answer to. What does it even mean not to use a protected attribute when you have lots of data and lots of correlations. Also, we were talking about families before. When we were doing our analysis e.g. how would algorithmic decisions compare to that of a human judge, the algorithm would release more black people, jail more Hispanics and jail more whites as well. Then, you can ask what if we release the same proportions of the subpopulations as the judges are releasing, and we still do better than the judge. But if we step back and ask what would be the right thing to do, we honestly don't know what should be the ratios. I think that's a big challenge - how do we think about this problem. The last anecdote that I will leave you is this. We did an experiment to understand where or why humans may be making mistakes. So we trained an algorithm that was trying to imitate the judge. So the algorithm didn't care about what's the right decision; it was just trying to imitate the judge. And when we took this artificial judge and applied it and saw what is it's accuracy, how good are the decisions that it made, this artificial judge was better than the judge it was trained on. And the only way to explain this is to say that the human judge has certain signals that the machine doesn't have access to. And whenever the machine makes a mistake - not imitating the teacher - the machine in some sense is making the correct decision. The machine didn't have access to certain signals that the judge was using in decision-making. The features that we were using in this work were based on history of criminal record - what was the age or sex, did they ever fail to appear before, and so on and so forth. These were features that were administrative, impossible to manipulate and the only way to affect them was to not commit

crime or to not get arrested I guess. There is the failure to appear, but there are also violent and non-violent crimes. We managed to reduce the violent crimes quite a bit. There is a good case why algorithmic decisions could help judges. The human judge can only see so many cases, the algorithm can see millions. When we were talking to judges, they told us that they have 30 seconds to a minute to make a decision. And after they make a decision, it is nearly impossible for them to see the outcome of the decision. They told us that the only way for them to learn whether it was a correct decision was to check the local newspapers and see if the released committed any crimes or not. So it's a very, very hard problem for the judges.

Jeanna Matthews: So the best criminals may not get arrested or may be people lucky enough to not live in jurisdictions with a high rate of arrest for crimes. Now, there are two points I would like to make. One is about the accuracy of data. We all know that in this world of big data, there is a lot of messy inaccurate data. So that's another important aspect of explanation. I cannot tell you how often I have looked at summaries like mean of the data and thought that is absolutely not true. I think there was a case recently about a proprietary recidivism software where somebody was saying one of the input pieces was incorrect, and they were arguing that it needed to be repeated. So it is not just consumer data but also your go-to-court kind of data where there are inaccuracies. So that's one thing. The other thing that I want to say is what are the forcing functions for debugging. I had a chance to go visit the Legal Aid Society of New York and they were talking about forensic software that are used to perform DNA matches of their clients. There was a software package where they got access to the source code, and which is now on Github. They found some very weird examples of bugs. One was where it should have been the case that the code erred on the side of not matching people and they found bugs where that was not the case. Also, that set of companies fight tooth and nail not to have disclosure of their software in any way in court, not even in a protected way, like not even the legal counsel gets to see it. Individual defense teams have to fight to get access to analyze the software. That's kind of a crazy world to be living in. It's very difficult to get the right to do that. And when you get in and see it, you find it's not doing what it's supposed to be doing. Maybe the data is inaccurate or messy. And in this world of criminal justice, maybe someone says I am not guilty, I swear I am not. And the system says, of course you say you are not, but you are a match and are going to jail. Again, I would like to ask what's the forcing function for debugging. Some of the defendants are true when they say they didn't do it. Are we going to lump all that together? If you have ever used a random software package, you know there are tons of bugs. You know there are bugs in there, right? What if you had to live with them forever because every time someone tried to report a bug, it was just dismissed? More importantly, what is the incentive of these companies to debug, to improve or make things better. They might feel that their software is perfectly fine. There might even be some buyers who are happier if it notches up the guilty ratings. They might be perfectly happy with the system as it is. They don't need any debugging, and don't need more accuracy or testing. There are some people going to jail. Our constituents are happy with that. We are good here... Until it is you or your family or friends. And also what population demographic is it more likely to be? There are just a lot of issues there.

Cynthia Dwork: I find this absolutely fascinating. I have a question. In cases where a mistake was found, is it something that required examination of the software or is it something where you already knew the answer and you were checking what the system output would be?

Jeanna Matthews: The specific case they were talking about involved source code analysis and finding a routine that did something which the software creators swore it did not do. I get why these companies might not want to reveal their software. But one of the more dangerous cases for me would be a company that said you can look at our system, it's completely open, but the problem is in the training data. Let me be a little more organized in my thought. One, you might have a problem in the data, not in the software at all. Two, companies don't want to reveal their software, so maybe we don't have to fight that battle.

Maybe what's better is targeted testing, being able to tweak things and see how the output changes. And some of the companies do provide some of these features. You could change little things and see what the answer would be. In the judicial context, they would prefer that than letting people look into the software. But then you would have to trust them, the answers they give back, this is the state of the system. And again the question I would ask is: what is the forcing function for them to make the system better and better and truly find bugs when they are just as happy without the extra effort? We all want to think that the systems we build are perfect and good and don't have any bugs. But apart from that what is the forcing function to find bugs, especially the harder corner case bugs that escape even source code examination.

John Langford: So with respect to such software, I think there should just be a law promoting transparency and open-source software. I don't see how we can trust a black-box to handle each case correctly.

Jeanna Matthews: I agree on open source for public use software. It's just that there are intellectual property rights issues that prevent open sourcing all proprietary software...

John Langford: But there is much more to a system than the algorithm. The algorithm can be made public and examined without all the system details.

Cynthia Dwork: Do you feel the same way about medical diagnostic devices that have circuits in them?

John Langford: I might...

Cynthia Dwork: It's interesting that we don't hear so much discussion of it.

John Langford: So another good example of bad data was Senator Patrick Leahy who discovered fake comments attributed to him that were anti-net neutrality even though he is pro-net neutrality. So there's a lot of bad data there.

The panel discussion turned at this point to answering audience question before finally summarizing the key takeaways from the discussion. The takeaways from each of the panelists are recorded below.

Jeanna Matthews: I will just reiterate that I think in some ways we might be changing our societal tradeoffs without any discussion by replacing some of our current processing with black-box decisions. That is something we should educate people about and care about. The potential for mischief in black-box systems is very high. I think we want to debug our systems, but not everyone who builds these systems may want to debug them and share our goals of transparency. If we begin to accept black-box decisions as being better than human decisions, that is a very dangerous road to go down on. Even if there is a cost in terms of accuracy, if we are talking about regulatable decisions, it's important to insist on explanations because the potential for mischief and bugs is too high, and the history of that kind of stuff is not good.

John Langford: I think we need a wider debate with society. I think there are two characteristics that make our current black-box decisions prone to bias: first is that they are black-box and second is that they are currently untestable. Sometimes, black-box systems are testable and that is enough, but if it's both black-box and untestable, then it's just ridiculous.

Jure Leskovec: My view would be that computer scientists or machine learning people or data people should actually go out there and be part of the debate and do real work. We can keep talking about this in our immediate community and write our papers, but the value of this is limited. When we get out of this conference zone and work on our concrete problem on a concrete application, people will care about that. This way I think we will learn much more about problems - what is a real problem and what is a made-up problem. We can then drive the agenda going further. What we learned in our research is that it's important to go out and say how can we do this better. Expose yourself, go out of our comfortable circles, and attack problems in the real world. This way new problems will arise, and we will solve them. We have to solve them, because none else will.

Cynthia Dwork: So I think it's a really good point. We have a lot of responsibility. Policy people don't understand the issues enough. When they are educated, my experience is that they turn around and say "ok, so what

do we do now." We just can't avoid trying to come up with answers. It's not that we have to get the final answers, but we certainly have to be able to discuss and bring wisdom to the conversation. On the lines of wisdom, we also need to be careful about the definition of terms. I think we can require companies to reveal their code, but you guys who know about theory of computing know that looking at code doesn't mean you have a clue what it is actually doing. There are fundamental questions that are still undecided. There is code obfuscation, and companies will exploit this if they don't want to reveal what they are doing. In the long run, I think we are going to have a situation where for example I am going to be represented by an artificial intelligence online which is going to go around and negotiate on my behalf, buy my airline tickets, etc. And this is another source I think of potential unfairness in the world. Take the artificial intelligence and replace it for example with a lawyer. People who can afford very good lawyers are going to win negotiations against people who can only afford much less expensive lawyers. And you can have a similar situation perhaps with artificial intelligences. The one that is going to represent me is perhaps not as good as the one that may represent a much richer person. And this is going to be exacerbated because things are going to happen really, really fast. So that's a whole another level of fairness that needs to be talked about.

At this point, Ricardo Baeza-Yates closed the panel discussion by thanking the panelists and the audience.

References

ACM US Public Policy Council. (2017). Statement on algorithmic transparency and accountability.

Abhinav Maurya is a graduate student at Carnegie Mellon University focusing on machine learning and public policy. He has previously studied at VJTI Bombay and IIT Bombay, and worked at Microsoft Redmond before heading to Pittsburgh to pursue doctoral studies.